



RESEARCH PROGRAM ON
Climate Change,
Agriculture and
Food Security



Creating and Using a Data and Document Storage Facility (DDS)

March 2018



Stats4SD

This document was produced in collaboration with [Statistics for Sustainable Development](#) and is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License](#).

Introduction

A Data and Document Storage Facility (DDS) is basically an area where project data and documents are stored, together with the rules that enable the project team to use it effectively and efficiently. For a single researcher working on his/her own, this could simply be a set of folders on his/her own PC hard drive. For a team of researchers this might be shared folders on a network file server or a web-based cloud storage system.

History

Before the advent of the Internet and computer networks, researchers would store their data and documents on standalone PCs. This worked fine for small projects with one researcher, but as projects became larger and teams developed, you often had the situation with each team member holding some but not all of the documents and data for the project. You often ended up with duplication and various copies of files with no one really knowing which was the most up to date. In short, there was a mess and archiving at the end of the project was just too daunting a task so wasn't done.

DDS vs Long Term Storage (Archives)

In these guides, we distinguish between 2 types of storage. A DDS is intended for active projects – where team members need regular access to data, and documents get updated regularly. Archiving occurs at the end of the project, when regular access is no longer needed, and files are not going to change much.

This may seem arbitrary, but the requirements of these 2 types of data storage are actually quite different, so it's likely you'll use different systems for them.

Why might I need a DDS?

The fact is, you always need a DDS for your project. Whether you are working alone or in a group, you need somewhere to store your project documents and data while working on them. The main questions are about how you organise this space, and whether you need a centralised space for your team to share resources.

When working alone, it's very tempting to not put effort into maintaining an organised system. We are all guilty at times of this: giving your vital documents a 'temporary' name; dragging it to your desktop for quick access and then forgetting to file it at the end of the day. Even when you can remember the file names and locations of things in the short term, this method will eventually fall down when you realise you need a specific document 6 months later and cannot remember if your most recent analysis script was "analysis working.R" or "analysis with metadata.R". (And that's assuming you know what project the files were for!)

When working in a team, good file organisation goes from important to vital. This is where you should consider a central storage space. Instead of the data and documentation being divided among the researchers to manage individually, you create a space to ensure the data and documentation relating to the project are up to date, complete and available to those who should have access. You should agree as a team how to use the space, and these agreements should be part of your Data Management Plan.

Having a central space doesn't stop a researcher taking a copy of part of the data to work on separately, as long as they know that any updates must be made to the master files in the DDS. The responsibility for negotiating and setting the rules for the DDS falls to the person with overall responsibility for the project, generally the PI. The responsibility for implementing the DDS is often delegated to the data manager.

Requirements of a good DDS

You can use many different systems to create your DDS, which we discuss below. Regardless of what tools you use, you should consider the following functions and how important they are for your project.

Storage

The primary use of a DDS is to store your active project files. This means you need to ensure you have enough space to store everything. When planning, try to predict how much storage you will need, and then set up your DDS with more than you think you need. That might mean buying more space on your cloud service account or buying extra drives for your local storage system. In 2018, storage is relatively cheap, so buying more than you think you'll need won't be a huge drain on your budget.

Beyond judgement of storage space, you may also want to consider the types of data and documents you'll be storing, and how they'll be used during the project. For example, if you are collecting large images and videos, you may want to provide local storage for people working with those files, as accessing them remotely might put strain on your network.

Syncing

Cloud storage systems like OneDrive, Dropbox etc. usually allow users to synchronise files to their local computers. This is hugely beneficial as it allows users to access files while offline and more easily work with locally installed applications.

Some systems allow for users to choose what files get synchronised, so they can have greater control over what files get stored on their local drives. This is increasingly important as many people now use laptops with relatively small solid-state drives.

Permissions Levels

When structuring your DDS, consider the purpose of each section, and who needs access. It is good practice to only grant specific types of access to those that need it. All good systems should let you choose between "read-only" and "read-write" for users. Beyond that, you might have other options, for example giving users admin permissions or the ability to grant access to other users. For example, when creating a shared folder in OneDrive, you can let users edit (read-write) or view (read-only). Someone with read-only access will not be able to edit the document directly within your DDS. They might be able to copy it somewhere else to edit, but your original file will remain intact and unchanged.

Version control

Many people use a “DIY” system of version control, by saving different versions of a file with slightly different names (see “Naming conventions” above). This is hugely preferable to having just a single file that you save over each time, as it makes you less vulnerable to accidental file corruption. If you choose to use this simple method of version control, we recommend making a copy of any file **before** starting your work for the day. That way, there is no way of accidentally overwriting your file with changes that you might later want to reverse.

This simple version control is usually enough for non-vital documents and personal files, but it relies on the user manually making copies to be a good option for mission-critical documentation or vital / unique datasets. For these more important files, we recommend a more automated option.

Certain apps have a form of version control or change tracking built in. Microsoft Office applications have a “track changes” feature that lets you follow the changes different authors make to a document. Google Docs also lets you review all the changes made by every author in a timeline, and lets you restore previous versions of a document.

But your DDS system can also include version control. For example, the Dropbox business plan lets you view and restore every version of a file saved within the last 120 days. OneDrive keeps file versions for 30 days, and OneDrive for Business lets you specify a certain number of versions to keep, rather than it being based on time.

A DDS kept on a local network drive may also have the capacity to keep different versions of files. Even on a local DDS, you might be able to setup some form of automatic version control. Windows 10 has a “File History” feature, which allows you to backup versions of your files in specific folders to an external drive, then access these older versions directly within the File Explorer.

Advanced Version Control – Git, SVN, Mercurial

There are more advanced version control systems focused on code development. Git, Subversion (SVN) and Mercurial are all systems focused around providing version control and change tracking for coding projects. They are extremely useful for software development and recommended for any project involving the writing of a lot of text files, for example doing analysis work or software development.

These systems take more time to learn and setup than a OneDrive folder or shared network drive but can provide huge benefits for certain types of work. If you want to be able to track line-by-line changes to code files, have a complete revision history (including comments) kept for you as you work, or allow different team members to work concurrently on the same codebase, we highly recommend exploring one of these systems.

Even if you’re not tracking text files, a system like Git might still work for you. There are projects that use Git to version control Word documents, Excel files and other large files (like repositories of images).

Backups

This is important: version control is not a backup. Using OneDrive as your DDS and syncing your OneDrive folder to your local computer does not count as having 2 copies of those files, because if

your local copy changes, the copy on the OneDrive server also changes. If someone accidentally deletes it in one place, it's gone everywhere. While you might be able to use OneDrive's version control to restore that file, that relies on a non-corrupt version being available in the system.

For this reason, your DDS plan should be more expansive than simply "We'll use OneDrive" or "We'll have a shared folder on the local network". You should consider what backup processes need to be implemented to ensure your data are safe. It may be that your DDS is already covered by an existing process. For example, your entire local network storage might be backed up every night by the network administrator, which would include a snapshot of your DDS. In this case, make sure you understand where those backups are kept, how secure they are and what processes are in place to access them if needed.

In general, you should consider the "3-2-1" rule of backups: Keep at least 3 copies of your data; 2 of which are local but on different devices, and at least 1 copy offsite. (If you are restricted from truly "offsite" storage by data policies or network restrictions, then try to spread the physical media out as far as possible – ideally in different buildings). This policy makes it extremely unlikely that you'll lose all copies of your data in one go.

A standard "manual" backup process might involve taking daily snapshots of your DDS and placing these, as compressed (.zip or .tar.gz files) on a secure storage – e.g. an external drive. To save space, you might only keep daily backups for 2 weeks. You might then keep weekly backups for up to 2 months, then beyond 2 months you might keep 1 backup for each month until the end of the project.

In practice, we recommend making the backup process as automatic as possible. The less you must rely on a member of your team manually copying files, the better. On a local Windows system, you might use the Windows Task Scheduler to run a batch file that clones your project folders to an external drive. Or you might choose to use a third-party backup tool. On a network, you might use rsync to run automatic, incremental backups of your network drive.

Security

You need to ensure your DDS is secure from unwanted access. If you are using local network storage, this is the responsibility of your network administrators. If you are using a third-party cloud storage solution, then it's the responsibility of your service providers, and you should check their documentation to ensure their security meets your needs.

Organising a DDS

One of the problems with a DDS is how to keep it well organised. We all know how difficult it is to organise our own work and our own files – most of us at some point will have overflowing mailboxes, full in-trays and will have spent a good half hour or so searching for that file that we know is "somewhere"! Imagine how much more difficult it is to keep a shared resource organised, especially when several researchers are adding and editing files.

Folder Structure

It is likely that you will have a folder structure for storing your files. Some systems use tags instead of folders, allowing you to assign multiple tags to each file. While tags are a much more flexible system,

folders are more common and generally easier to manage. This is likely because it's much easier for us to imagine a file as a physical document, placed in a single place inside a folder system.










It's important to clearly structure your DDS, and this means having clear names for your folders. We recommend including a document at the root of your DDS that describes the folder structure, so everyone in the team is aware of what should go inside each folder.

Some people favour deeply nested folders, while others prefer shallow folder structures. Ultimately, it doesn't really matter how you organise your files, as long as you can effectively communicate the structure to your team and everyone agrees on where files should go.

Naming Conventions for Files in the DDS

We suggest you develop a naming convention for files within the DDS. One method that is often used is to include the date within the file name. For example, Figure 1 shows a list of files where the name always starts with the date. The date here is in the form YYYY-MM-DD. We recommend this format to make it easy to sort – here, sorting by filename puts the files in date order. You can see that there are two documents called “Dataverse.docx” but one includes the date as 2017-12-07 and the other has the date as 2017-12-11. Don't rely on the modification date as this often picks up the date a file was moved or copied; if you have any Access databases the date modified will change whenever the database was opened regardless of whether or not any changes were made.

Figure 1 - Using Date in the file name

| Name | Date modified | Type | Size |
|---|------------------|---------------------|----------|
|  2017-12-07 Dataverse.docx | 07/12/2017 09:03 | Microsoft Word D... | 13 KB |
|  2017-12-07 DDS.docx | 07/12/2017 10:28 | Microsoft Word D... | 14 KB |
|  2017-12-07 Metadata.docx | 07/12/2017 10:49 | Microsoft Word D... | 13 KB |
|  2017-12-07 Tools for Research Projects.docx | 07/12/2017 10:41 | Microsoft Word D... | 14 KB |
|  2017-12-07 Training Manual Example using ODK.docx | 07/12/2017 14:50 | Microsoft Word D... | 1,500 KB |
|  2017-12-08 Introduction to Dataverse.docx | 11/12/2017 09:46 | Microsoft Word D... | 256 KB |
|  2017-12-11 Data and Document Storage.docx | 11/12/2017 10:03 | Microsoft Word D... | 256 KB |
|  2017-12-11 Dataverse.docx | 11/12/2017 10:07 | Microsoft Word D... | 14 KB |
|  2017-12-11 Introduction to Dataverse.docx | 11/12/2017 10:04 | Microsoft Word D... | 256 KB |

Such naming conventions may seem either obvious or unnecessary, but not defining them at the start can lead to a mess later on. Figure 2 shows a number of files stored together in the same folder. These files are actually different versions of the same document, but it would be very different to find the “correct” version or understand what the differences are between the files.

Figure 2 - Versions of the same document

| | | |
|---------------------------------------|------------------|---------------------------------|
| pr-fig1.u.shg | 01/02/2000 16:14 | SHG File |
| p-tgs-2000-10.doc | 25/10/2000 17:50 | Microsoft Office Word 97 - 2003 |
| p-tgs-2000-05.doc | 18/05/2000 17:00 | Microsoft Office Word 97 - 2003 |
| p-tgs-99.doc | 18/10/1999 13:34 | Microsoft Office Word 97 - 2003 |
| p-tgs-00.doc | 23/03/2000 11:45 | Microsoft Office Word 97 - 2003 |
| p-tgs.doc | 16/05/2001 11:59 | Microsoft Office Word 97 - 2003 |
| PresentingResults.doc | 21/01/2000 12:40 | Microsoft Office Word 97 - 2003 |
| presenting results- revised by RC.doc | 15/11/1999 08:54 | Microsoft Office Word 97 - 2003 |
| Presentation.doc | 04/04/2000 15:07 | Microsoft Office Word 97 - 2003 |
| Presentation booklet.doc | 11/01/2000 12:14 | Microsoft Office Word 97 - 2003 |
| present.doc | 03/02/1998 13:38 | Microsoft Office Word 97 - 2003 |
| h-tgs.doc | 16/05/2001 12:02 | Microsoft Office Word 97 - 2003 |
| h-gfp.doc | 15/04/1998 08:47 | Microsoft Office Word 97 - 2003 |
| penhouse.tif | 12/12/2002 18:02 | TIFF File |

Sorting out the mess

Even with an agreed naming convention, situations like in figure 2 are common. If you manage to catch a mess before it gets too big, you might be able to organise the files by talking to your team and figuring out what the files should be called. But that's not always possible, and a big mess might need a lot of time to sort out.

One suggestion is to create a 'backlog'. Move the mess into a separate space dedicated to your backlog. Then, if you haven't already got a system in place, set up your DDS as if you were starting from scratch. Any new documents and data can go into the new clean structure with names following the agreed convention. Doing this means that your backlog doesn't hold up new work, and you can spend a bit of time each day or week sorting it out. Ideally, you will eventually get through your backlog and have it all fully organised into your main system.

Data Storage Models

Many data storage models can be used for the DDS. A decentralised system where every person on the team keeps files but each person deposits files into a single place for safekeeping and as a depository of the most up to date version of all data and documents could be used. This has the advantage of freedom but the major disadvantage that ensuring completeness and up-to-date documents is very difficult.

With current technology a centralised system is quite appealing because it makes it easy to ensure completeness and up-to-date documents without much need for coordination of people. One disadvantage to a central store to which everyone has access is that it can easily become a dump which is then very difficult and time-consuming to sort out. To avoid a "dump" you might consider appointing someone as the "custodian" of the DDS. Data and documents go to and from the DDS via this person. An alternative would be to give everyone read access but only give write access to the custodian. A compromise solution is to have a shared development folder where everyone has read/write access but also a folder which has read-only access where final versions of files are stored.

Ownership of the DDS

There is a distinction between the ownership of the DDS and the ownership of the data and documents stored within. A DDS system might be stored on OneDrive, Dropbox, Google Drive or an internal system from your organisation. This does not mean that all the data stored within it belong to those institutions.

Many researchers still have issues over sharing “their” data, somehow believing that in doing so they are giving away their rights and someone else will get the credit for their work. However, you should bear in mind that ownership is generally defined by a contract between the funding agency and the researcher or organisation. Check your contract to ensure you know what this is. Also, it is a good idea to draw up agreements that can be signed by all members of the project team. This ensures that everyone knows where they stand with respect to data ownership. Remember that the Intellectual Property Rights (IPR) of the contents of the DDS remain the same regardless of where the DDS is stored and who has access to it. Ownership does not change simply because you have placed the files into a shared location.

Ultimately, a shared DDS requires some degree of trust. Remember, though, this works both ways: if you are not willing to share your data and documentation with others, you cannot expect them to share their files with you.

What System(s) Should you use?

Where you store your DDS depends on the resources and local skills you have available.

A Cloud Storage Service

If you have good internet access, an easy solution is to use a third-party cloud storage solution. As of 2018, the three main contenders are:

- OneDrive / SharePoint
- Dropbox
- Google Drive.

Each company offers a variety of plans, including ones designed for business use. There are a host of other options too and each service offers similar features.

Benefits

Benefits of a cloud storage system include:

- You can access your DDS anywhere you have an internet connection, making it useful for remote teams;
- Responsibility for server maintenance is passed to the service provider;
- A good service provider has the benefit of scale – they manage far more storage space than your project, so can provide a much cheaper and more stable system than a custom-purpose server.
- Your master files are not stored on any one person’s computer, so you are protected against hardware failures.

Issues

A common concern about cloud storage options is that of security and privacy. Using any internet-based system does increase your security risks slightly, simply by the fact of your data being accessible from more locations. All good systems offer fully encrypted data transfers and encrypted storage. For example, Dropbox uses SSL / TLS to encrypt all data in transit and encrypts data on its servers with 256-bit AES – an industry recognised standard. We recommend carefully reading the documentation for the service you are considering to ensure it meets the requirements of your project.

Organisations may have policies indicating what can and cannot be placed onto a third-party system. For example, a university might require that any personal data about human subjects not be placed onto a third-party system to comply with data protection laws.

A Shared Network Drive

If your team are in the same location, or can access a local network via a VPN, you could use a shared network drive to which all team members have access.

This is a good solution if you are required to keep data and documents “on-premises”, or do not have a stable internet connection. It means that your team (or someone in your organisation) has responsibility to manage the infrastructure.

With the current pace of change in technological development new solutions are appearing all the time. The important thing is not which technology to use but rather to be aware that establishing a DDS is highly beneficial for a research project and making the managerial decisions that will make use of the best technology available to help achieving good management of the data and documents of a research project.

Summary

The Data and Document Store is a system to help you keep all your project files together in a centralised location. A well-organised DDS means that team members can always access the latest documents and data and data integrity is preserved. Archiving at the end of the project is made quicker and easier.

Remember though, there is no special software involved and there is certainly no magic wand to organise your files. As a team you must decide on the structure of your DDS to ensure it becomes a useful resource and not just a file dump.

Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at <https://www.youtube.com/channel/UCs7EU95YMjHvNozJKCD92xQ/playlists>. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes the following videos associated with Data and Document Storage:

Introduction to Data and Document Storage -

https://www.youtube.com/watch?v=4CQtJbg_Qms&index=6&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpi

Ownership Issues with Data and Document Stores -

<https://www.youtube.com/watch?v=ML3UXLzaqRw&index=8&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpi>

Data and Document Store Organisation -

https://www.youtube.com/watch?v=MMagU_77rdI&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpi&index=7